Mots audio-visuels joints pour la détection de scènes violentes dans les vidéos

Nadia Derbas, Georges Quénot

▶ **To cite this version:**

Nadia Derbas, Georges Quénot. Mots audio-visuels joints pour la détection de scènes violentes dans les vidéos. CORIA, 2014, Nancy, France. pp.63-77. hal-00981707

HAL Id: hal-00981707
https://hal.science/hal-00981707

Submitted on 22 Apr 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Mots audio-visuels joints pour la détection de scènes violentes dans les vidéos

Nadia Derbas — Georges Quénot

UJF-Grenoble 1 / UPMF-Grenoble 2 / Grenoble INP / CNRS, LIG UMR 5217, Grenoble, F-38041, France

{Prénom.Nom}@imag.fr

RÉSUMÉ. *Ce papier présente une représentation audio-visuelle des données pour la détection des scènes violentes dans les films. Les travaux existants dans ce domaine considèrent l'information visuelle ou l'information audio; voire leur fusion classique. Jusqu'à présent peu d'approches ont exploré leur dépendance mutuelle pour la détection de scènes violentes. Ainsi, nous proposons un descripteur qui fournit des indices multimodaux audio et visuels; tout d'abord en assemblant les descripteurs audio et visuels, ensuite en révélant statistiquement les motifs conjoints multimodaux. La validation expérimentale a été effectuée dans le cadre de la tâche « détection de scènes violentes » de MediaEval 2013. Les résultats obtenus montrent le potentiel de l'approche proposée en comparaison avec les méthodes utilisant les descripteurs audio et visuels séparément ou d'autres types de fusion.*

ABSTRACT. *This paper presents an audio-visual data representation for violent scenes detection in Hollywood movies. Existing works in this field consider either the audio or the visual information; or their shallow fusion. None has yet explored their joint dependence for violent scenes detection. We propose a feature which provides strong multimodal audio and visual cues by first joining the audio and the visual features and then revealing statistically the joint multimodal patterns. Experimental validation was conducted in the context of the "Violent Scenes Detection" task of the MediaEval 2013 Multimedia benchmark. The obtained results show the potential of the proposed approach in comparison to methods using audio and visual feature separately and other fusion methods.*

MOTS-CLÉS : *Indéxation sémantique, Analyse de contenu, Fusion audio-visuelle, Multimédia, MediaEval.*

KEYWORDS: *Semantic Indexing, Content Analysis, Audio Visual Fusion, Multimedia, MediaEval.*

1. Introduction

Des milliers de films sont produits par l'industrie cinématographique chaque année. Devant cette impressionnante quantité de documents, le besoin de techniques de recherche d'information dans les documents multimédia se fait sentir afin de pouvoir classifier automatiquement ces films. Plusieurs techniques d'analyse automatique de films ont été mises en place et implémentées, comme la classification par catégorie (drame, fiction, comédie, ...) ou encore la caractérisation de scènes. Il a été souligné qu'un grand nombre de films ne sont pas adressés au grand public et auraient des effets nocifs sur la santé psychologique des spectateurs. C'est notamment le cas des films violents lorsqu'ils sont regardés par un public très jeune ou des enfants. Pour cela, des approches ont déjà été proposées pour résoudre le problème de la détection automatique de violence dans les films de manière à empêcher que des enfants puissent visionner de telles images.

Définir le terme « violence » n'est pas une tâche facile à cause de son ambiguïté et de sa subjectivité. Certains spectateurs pourront voir de la violence où d'autres n'en verront pas. Dans le domaine de recherche, chaque scientifique a dû clarifier sa propre description de la violence. On peut trouver des définitions littéraires comme : « violence physique ou accident amenant à des blessures humaines ou de la douleur » (Demarty *et al.*, 2013). Il existe aussi des définitions plus techniques où la violence est définie par des indicateurs visuels et audio spécifiques, par exemple les mouvements accélérés ou les rythmes de musique rapides (Gong *et al.*, 2008).

Dans ce papier, nous proposons une méthode efficace pour la détection automatique de scènes violentes dans le contexte de films hollywoodiens. Notre approche est basée sur une représentation sous la forme des sacs-de-mots audio-visuels. Cette représentation décrit le contenu de la vidéo en se basant sur la relation conjointe entre les deux modalités en même temps : la modalité audio et la modalité visuelle. Le reste du papier est organisée de la façon suivante : dans la section 2 nous présenterons les travaux connexes à notre travail. Nous décrirons notre méthode et ses différentes étapes dans la section 3. Dans la section 4, nous évaluerons la méthode proposée sur la collection de données de MediaEval 2013 et nous commenterons les résultats obtenus. Enfin, dans la section 5 nous tirerons nos conclusions.

2. Travaux Connexes

L'état de l'art concernant la détection de contenu violent est limité et souvent basé sur des descripteurs visuels ou spatio-temporels (Datta *et al.*, 2002 ; Bermejo Nievas *et al.*, 2011 ; de Souza *et al.*, 2010). D'autres méthodes se focalisent sur l'unique utilisation des descripteurs audio (Giannakopoulos *et al.*, 2006). Certains proposent même une nouvelle utilisation de la représentation en sacs-de-mots audio classiques, en décrivant chaque segment par un ou plusieurs mots audio obtenu par une simple

agrégation sur les descripteurs audio classiques (Penet *et al.*, 2013).

Les scènes violentes dans les films apparaissent souvent avec des événements audio spécifiques (par exemple des coups de feu) et forment donc des motifs audio-visuels. Ainsi nous pensons qu'une méthode efficace pour la détection de scènes violentes doit exploiter les deux modalités : audio et visuelle. La technique la plus populaire d'analyse audio-visuelle est basée sur la fusion multimodale. Altrey *et al.* ont étudié un grand nombre de méthodes de fusions multimodales proposées durant ces dernières années, afin d'apporter une meilleure compréhension et un meilleur choix en fonction de la problématique (Atrey *et al.*, 2010). Nous pouvons distinguer trois principaux types de fusion :

– **la fusion précoce** : dans la fusion précoce le descripteur audio et le descripteur visuel sont combinés avant la classification (Gong *et al.*, 2008). La simplicité d'implémentation de cette fusion (concaténation des descripteurs des différentes modalités) l'a rendu populaire dans le domaine. Cependant, une grande différence d'échelle des valeurs ou du nombre de dimensions des descripteurs concaténés peut entrainer un déséquilibre dans la prise en compte de certains descripteurs par rapport aux autres. Dans ces cas, il sera nécessaire de normaliser ou de pondérer les différents descripteurs après la concaténation. Également, la concaténation de plusieurs descripteurs peut générer des descripteurs à très grandes dimensions et entrainer par la suite un temps d'apprentissage beaucoup plus long. Dans ces cas il sera nécessaire de réduire la dimension du descripteur final en appliquant une méthode de réduction de dimensions, par exemple une analyse en composantes principales (ACP).

– **la fusion tardive** (Snoek *et al.*, 2005) : dans la fusion tardive les scores de classification obtenus séparément par chacun des modèles de descripteur sont combinés (Derbas *et al.*, 2012 ; Giannakopoulos *et al.*, 2010 ; Lin et Wang, 2009). Contrairement à la fusion précoce, la fusion tardive s'appuie sur la force de chacune des modalités séparément. L'avantage de cette fusion est la possibilité d'utiliser la méthode de classification la plus appropriée à chacune des modalités (celle qui considère au mieux la spécificité de la modalité). De plus, la combinaison de différentes méthodes de classification permet de palier les erreurs de prédiction de chacune séparément et fournit souvent des décisions plus précises. Néanmoins, ce gain de précision a un coût qui se traduit en une augmentation de temps de calcul vu que chaque modalité nécessite sa propre étape d'apprentissage.

– **la fusion de noyaux** : La fusion de noyaux peut être considérée comme une fusion intermédiaire entre la fusion précoce et la fusion tardive. La fusion à noyaux combine les modalités au niveau du noyau. Au lieu de modéliser les données selon une seule fonction de noyaux (comme dans la fusion précoce), cette fusion offre la possibilité de choisir le noyau le plus adapté à chacune des modalités et de combiner ensuite les noyaux uni-modaux pour générer un seul noyau final multi-modal

(Ayache *et al.*, 2007 ; Mühling *et al.*, 2012). Elle permet d'exploiter le maximum d'informations de chacune des modalités. L'inconvénient de cette méthode de fusion est le nombre de paramètres à fixer d'abord sur l'ensemble des fonctions de noyaux pour chacune des modalités, et ensuite sur la fonction de fusion pour le noyau final.

Le principal problème de ces méthodes est qu'elles fusionnent les modalités audio et visuelles sans prendre en considération leurs corrélations potentielles. Les approches de fusion précoce de l'état de l'art représentent séparément le contenu des vidéos en concaténant n histogrammes monodimensionnels correspondant aux n modalités considérés. Nous proposons dans ce papier, une approche qui représente le contenu des vidéos conjointement par l'intermédiaire d'un seul histogramme multidimensionnel (à n dimensions) afin de mieux prendre en compte la corrélations entre les différents modalités. En effet, les histogrammes multidimensionnels fournissent, habituellement, une représentation plus fine du contenu que celle de plusieurs histogrammes monodimensionnels. Par exemple dans le cas des images, l'histogramme tridimensionnel de couleur RGB est plus distinctif que trois histogrammes monodimensionnelles (R, G et B). Pour illustrer cette différence, nous considérons deux images dont la première contient du rouge et du bleu et dont la deuxième contient du noir et du violet. Bien que ces deux images soient visuellement très différentes, les trois histogrammes monodimensionnnels (R, G et B) en donnenent exactement la même représentation. L'histogramme tridimensionnel RGB est par contre en mesure d'en fournir deux représentations effectivement très différentes. La Figure 1 illustre cet exemple. Notons que les méthodes de fusion par noyaux et tardives prennent encore moins en compte la corrélation entre les éléments des différentes modalités puisque la fusion se fait encore plus loin du signal.

Cependant, nous avons trouvé dans des domaines connexes quelques travaux intéressant qui analysent conjointement le contenu audio et le contenu visuel des vidéos. Pour la détection des événements dans les vidéos, Ye *et al.* ont modélisé la relation entre la modalité audio et la modalité visuelle avec un graphe biparti suivie d'un partitionnement de ce graphe de façon à révéler des motifs joints (Ye *et al.*, 2012). Alors que dans le domaine de reconnaissance des événements dans les vidéos de camera de surveillance, certains ont proposé des méthodes pour intégrer les informations audio et visuelles (Cristani *et al.*, 2007). Ils calculent une matrice de co-occurrence audio-visuelle pour détecter et segmenter les événements sous la forme de données audio visuelles. Dans le domaine du suivi d'objets, Beal *et al.* ont décidé d'exploiter la structure statistique des données audio et visuelles ainsi que leur dépendances mutuelles. Ils l'ont traduit dans un seul modèle graphique probabiliste (Beal *et al.*, 2003). Enfin dans le domaine général de classification de concepts dans les vidéos, Jiang *et al.* ont étudié la causalité temporelle statistique entre les mots audio et visuels pour représenter le contenu de la vidéos comme étant des motifs audio-visuels (Jiang et Loui, 2011).

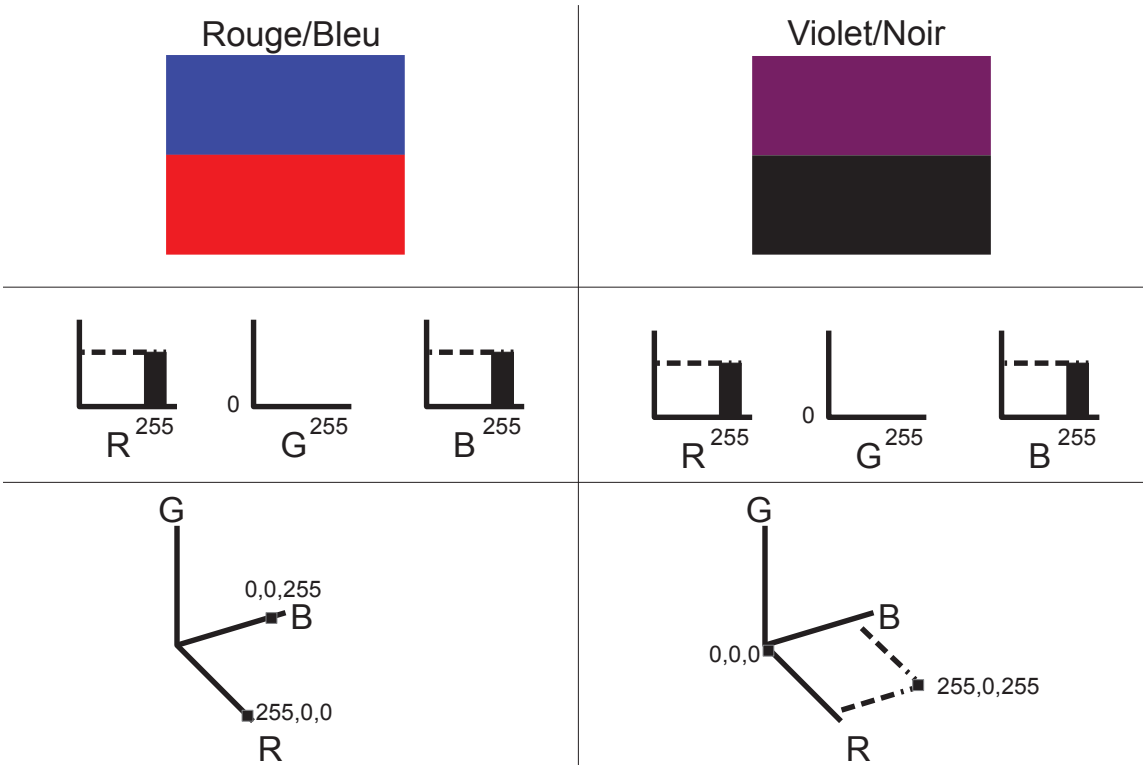

Figure 1 – Les représentations fournient par un histogramme tridimensionnel RGB versus celle fournit par trois histogrammes monodimensionnels (R, G et B) pour deux images distinctes.

3. Descripteur audio-visuel joint

3.1. *Extraction des mots audio-visuels joints*

Cette section décrit la représentation audio-visuelle jointe que nous proposons pour la détection de scènes violentes. Le but étant d'exploiter la forte corrélation entre l'information audio et l'information visuelle afin de découvrir des motifs audio-visuels capables d'identifier les scènes violentes. Les motifs audio-visuels sont censés donner de meilleurs résultats qu'une simple fusion (précoce ou tardive) des deux modalités audio et visuelle qui ignore leurs corrélations. La méthode proposée est composée de trois étapes :

1) En premier temps, les descripteurs locaux audio et visuels sont extraits à partir de la vidéo ;

2) Ensuite, les motifs bimodaux (ou encore les mots bimodaux) sont retrouvés et le dictionnaire bimodal est construit ;

3) Enfin, la représentation sous la forme de sacs-de-mots bimodaux est construite par l'intermédiaire des ces mots.

Le processus général de la méthode est illustré dans la Figure 2.

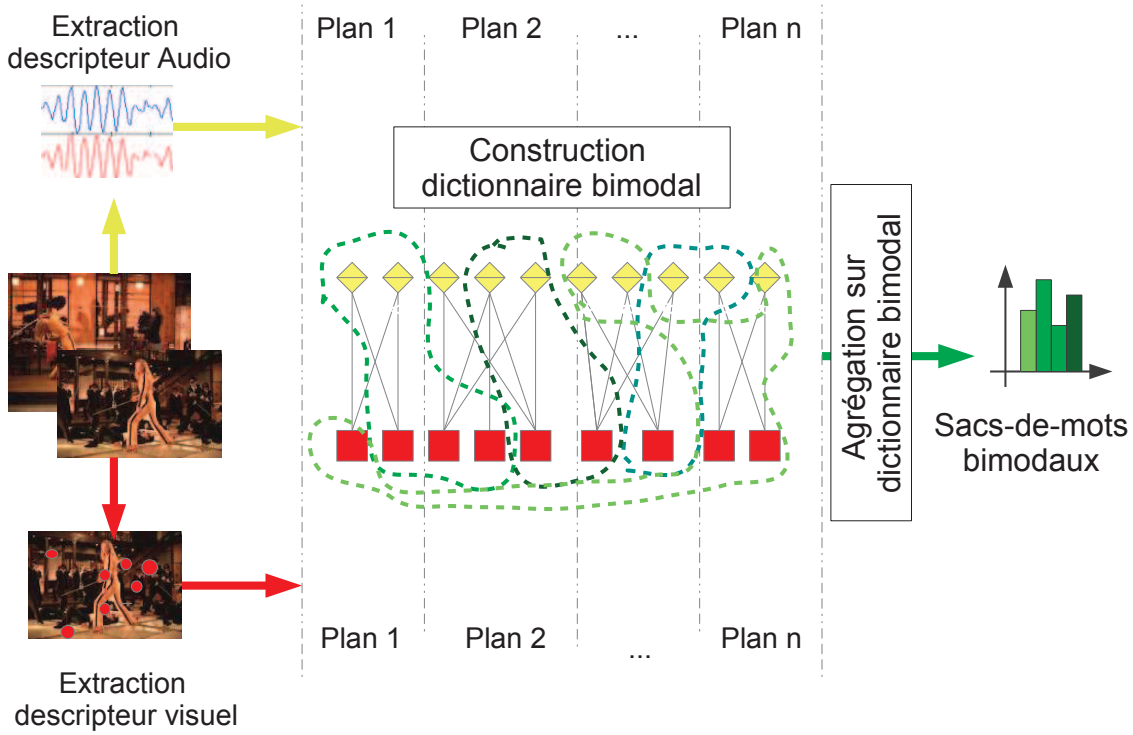

Figure 2 – Le processus général pour la génération de sacs-de-mots audio-visuels

3.1.1. *Extraction des descripteurs*

On considère une collection de vidéos décomposée en n_s plans vidéo. Les descripteurs locaux audio et visuels sont extraits pour chacun des plans. Pour chaque plan vidéo s_i, ceci peut être écrit sous la forme de a_{ij} et v_{ik} avec $1 \leq i \leq n_s$, $1 \leq j \leq n_{ai}$ et $1 \leq k \leq n_{vi}$, et avec n_{ai} et n_{vi} étant respectivement le nombre de descripteurs audio locaux et le nombre de descripteurs visuels locaux dans le plan vidéo s_i. Tous les a_{ij} (resp. v_{ik}) sont des vecteurs de dimensions fixées à d_a (resp. d_v) et le nombre de descripteurs locaux n_{ai} et n_{vi} dépend généralement du contenu du plan vidéo s_i.

3.1.2. *Capture de motifs bimodaux*

Dans l'approche classique de sac-de-mots et par modalité, a_{ij} (resp. v_{ik}) sont agrégés sous la forme d'histogramme selon les groupes (clusters) calculés précédemment sur la totalité des descripteurs locaux du même type disponibles sur la collection de données. Les groupes (clusters) constituent alors un dictionnaire de taille fixe prédéfinie qui correspond également à la taille de représentation agrégée.

Pour la représentation audio-visuelle jointe, on considère un ensemble de $m_{ijk} = a_{ij} \oplus v_{ik}$ vecteurs avec $1 \leq i \leq n_s$, $1 \leq j \leq n_{ai}$ et $1 \leq k \leq n_{vi}$, et où $a_{ij} \oplus v_{ik}$ est simplement la concaténation de a_{ij} et v_{ik}. Pour cela tous les m_{ijk}

doivent avoir une même dimension fixe $d_a + d_v$ et le nombre de descripteurs locaux $n_{ai} \times n_{vi}$ dépend généralement du contenu du plan vidéo s_i. Avant la concaténation des descripteurs audio et visuels, une normalisation et éventuellement une pondération peuvent être appliquées. La normalisation peut être faite par un coefficient multiplicatif global de façon à ce que la distance moyenne entre deux descripteurs locaux soit égale à 1 pour ramener les descripteurs à échelle équivalentes. En ce qui concerne la pondération, celle ci peut être effectuée selon la performance relative des descripteurs audio et visuels considérés séparément et évalués par une validation croisée sur l'ensemble d'apprentissage.

Le nombre de descripteurs audio peut être très élevé, de même pour le nombre de descripteurs visuels pour chacun des plans vidéo de la collection. De plus, le grand nombre de plans vidéos dans la collection de données peut entraîner la génération d'un très grand nombre de descripteurs audio-visuels joints. Même si la représentation n'aura pas à être stockée le processus d'agrégation doit lui être appliqué, d'où le besoin de trouver une solution pour réduire leur nombre. Dans le but de rendre l'approche généralisable et applicable dans le cas où plus de deux descripteurs locaux auront à être fusionnés de cette manière, nous proposons de limiter le nombre de combinaisons audio-visuelles considérées à une certaine valeur n_{max}. Dans ce cas, la représentation locale audio-visuelle jointe sera un ensemble de m_{il} avec $1 \leq i \leq n_s$ et $1 \leq l \leq n_{ml}$ avec n_{ml} étant le nombre de descripteurs locaux audio-visuels dans le plan vidéo s_i. Dans le cas où $n_{ai} \times n_{vi} \leq n_{max}$, on prend $n_{ml} = n_{ai} \times n_{vi}$ et tous les $a_{ij} \oplus v_{ik}$ sont considérés. Dans le cas où $n_{ai} \times n_{vi} > n_{max}$, on prend $n_{ml} = n_{max}$ et seulement n_{max} vecteurs sont sélectionnés aléatoirement $a_{ij} \oplus v_{ik}$ à partir de $n_{ai} \times n_{vi}$ sont considérés. Enfin une méthode standard de regroupement sera appliquée sur n_{max} vecteurs joints pour capturer la corrélation entre l'information audio et visuelle et donc retrouver les motifs audio-visuels.

3.1.3. *Représentation sous la forme de sacs-de-mots bimodaux*

Enfin, l'agrégation de type sacs-de-mots peut être appliquée sur l'ensemble d'apprentissage, exactement de la même manière sur les m_{il} descripteurs locaux que si elle a été appliquée sur a_{ij} et v_{ik}.

3.2. *Choix de paramètres*

Pour générer le descripteur audio-visuel joint proposé, nous utilisons un descripteur audio classique « Mel Frequency Cepstral Coefficients » (MFCC) pour représenter le contenu audio de la vidéo. Les points d'intérêt spatio-temporels (STIP) sont quant à eux utilisés pour représenter l'information visuelle de la vidéo (Laptev, 2005) vu que le mouvement est très important pour la détection de violence et que ce descripteur met l'accent sur le mouvement. Avant la génération de la représentation audio-visuelle jointe, les deux descripteurs ont été optimisés séparément sur les 12 concepts annotés (fournis par les organisateurs de la tâche de

détection de scènes violentes de MediaEval 2013) pour éviter le sur-apprentissage sur les deux concepts cibles (violence objective et violence subjective). La classification a été effectuée sur deux méthodes d'apprentissage différentes, une basée sur des SVM multiples (MSVM) (Safadi et Quénot, 2010) et une autre basée sur la recherche de K plus proches voisins.

L'outil d'Ivan Laptev a été utilisé pour calculer les points d'intérêt spatio-temporels (STIP) (Laptev, 2005). Un vecteur d'histogramme de flux optique (HOF) est ainsi produit pour chaque STIP détecté. La dimension de ce vecteur est de 90 éléments. Une agrégation est ensuite appliquée sur la durée du plan.

L'outil de Guillaume Gravier Spro[1] a été utilisé pour le calcul du descripteur audio MFCC. Un vecteur dont la dimension est de 10 éléments est produit par le programme chaque 10 ms. La durée minimale de la fenêtre impacte directement la performance de ce descripteur audio. Nous avons optimisé cette durée minimale par validation croisée sur l'ensemble d'apprentissage. La Figure 3 montre la performance des MFCC en fonction de différentes durées de fenêtre. Une durée minimale de 1.8 secondes donne les meilleures performances. Une agrégation est appliquée sur la durée du plan étendue avant et/ou après pour atteindre 1.8 secondes si celle-ci est inférieure à cette valeur.

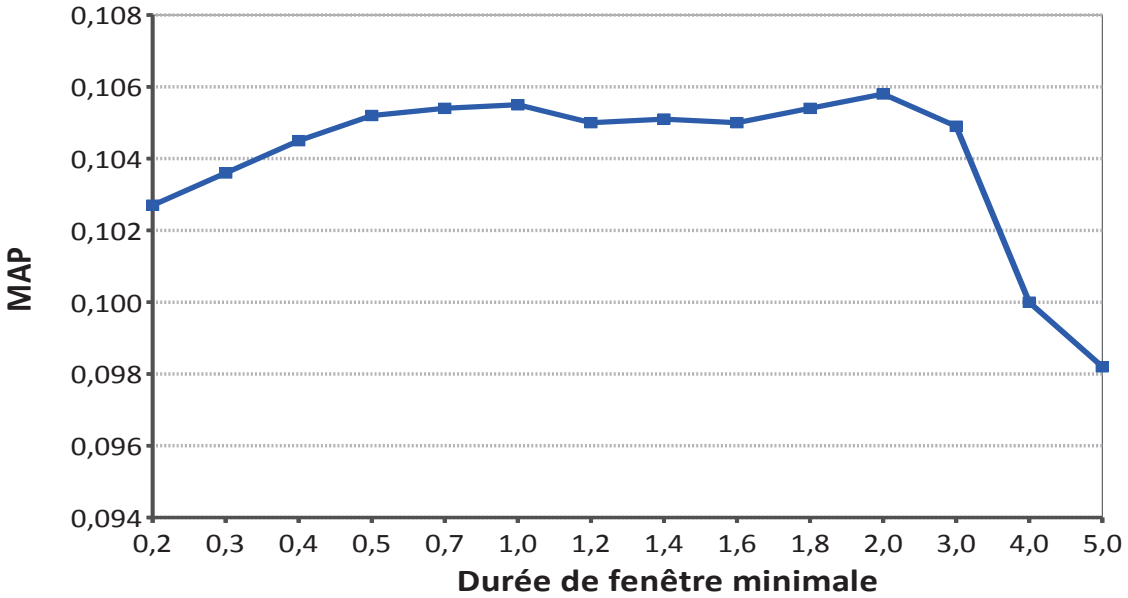

Figure 3 – La performance des MFCC en fonction de la durée minimale de la fenêtre.

1. http://www.irisa.fr/metiss/guig/spro/

De plus, le nombre de groupes (clusters) calculé sur les descripteurs locaux extraits (taille du dictionnaire) et utilisé pour la génération de la représentation en sacs-de-mots influence sensiblement la performance de ces descripteurs. Nous avons également optimisé la taille du dictionnaire par une validation croisée sur l'ensemble d'apprentissage. L'influence du nombre de groupes sur la performance du système global est illustrée par la Figure 4. Le nombre de 4096 groupes a donné les meilleurs résultats, donc toutes les agrégations par modalité ou pour le joint ont été calculées en utilisant une représentation en sacs-de-mots avec un dictionnaire de taille 4096.

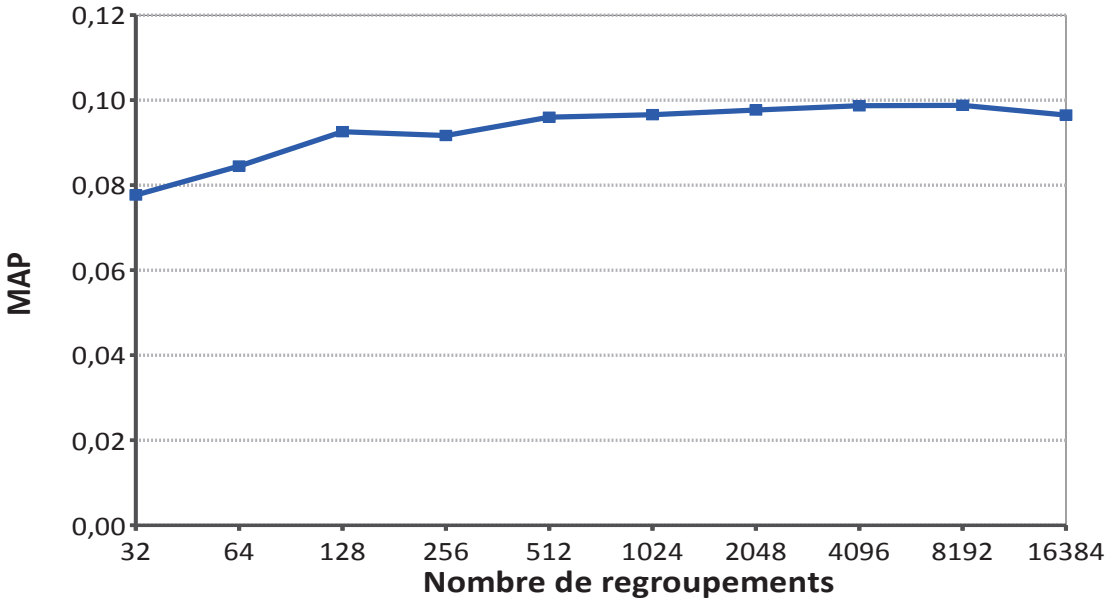

Figure 4 – La performance des MFCC en fonction du nombre de groupes.

Enfin, nous avons fixé le nombre de combinaisons des vecteurs audio-visuels considéré (n_{max}) expérimentalement par validation croisée sur l'ensemble d'apprentissage. Nous l'avons évalué avec différentes valeurs de n_{max}, pour des raisons de complexité et de temps de calcul nous nous sommes arrêtés à 32 768 combinaisons surtout que la courbe obtenue commençait à se stabiliser. Le meilleur résultat a été obtenu avec 32 768 comme le montre la Figure 5.

4. Expérimentations et résultats

4.1. *Collection de données*

L'efficacité de notre représentation audio-visuelle jointe a été mesurée dans le cadre de la tâche de détection de scènes violentes de MediaEval2013. Cette tâche

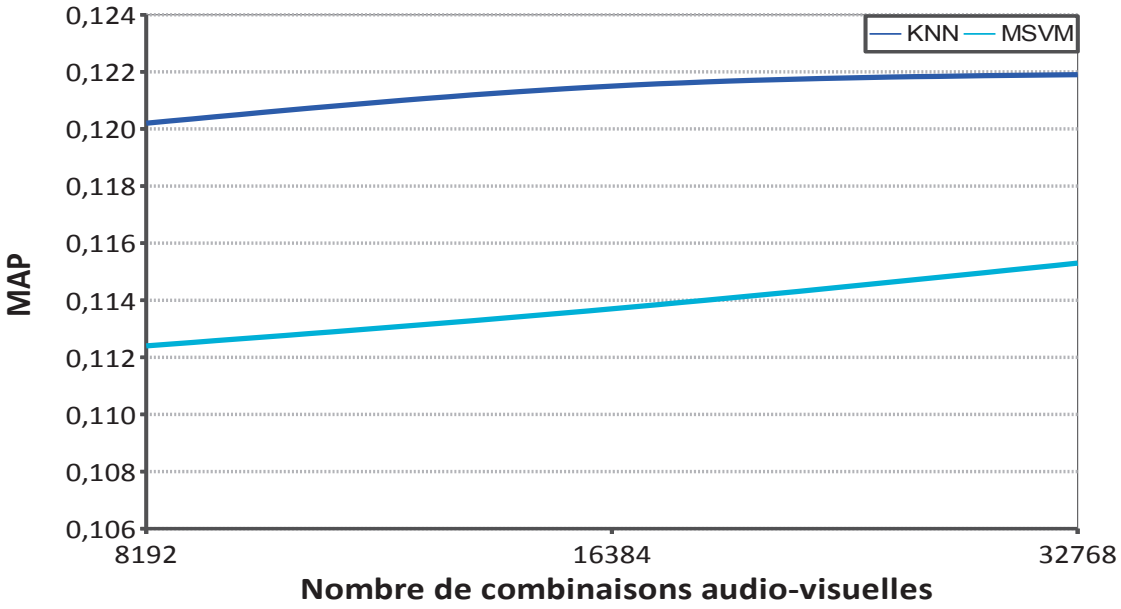

Figure 5 – L'influence du n_{max} sur la performance du descripteur audio-visuel joint.

définit deux types de violence : violence objective et violence subjective. La violence objective est définie comme étant « violence physique ou accident résultant en blessures humaines ou douleur ». La violence subjective est définie comme étant « les scènes qu'on ne pourra pas laisser un enfant de 8 ans regarder dans un film à cause de la violence physique qu'elles contiennent » (Demarty *et al.*, 2013). La Figure 6 montre quelques images extraites des plans annotés comme contenant des scènes violentes objectives.

Figure 6 – Quelques images extraites des plans annotés comme étant violents

La collection de données comprend 25 films hollywoodiens décomposés en 43 923 plans et répartis en deux ensembles : l'ensemble d'apprentissage et l'ensemble de test. L'ensemble d'apprentissage contient 18 films annotés : *Kill Bill, The sixth sense, Armageddon* ... décomposés en 32 678 plans. Alors que l'ensemble de test contient 7 autres films hollywoodiens décomposés en 11 245 plans. Les données de l'ensemble d'apprentissage sont annotées par plan comme contenant ou non des scènes violentes (subjective ou objective) en plus de dix autres concepts : *sang, feu, cris, poursuite de voitures, arme à feu, gore, arme blanche, explosions, coups de feu, batailles*. Ces dix concepts peuvent être utilisés par les participants pour détecter les scènes violentes. Le nombre de plans annotés comme étant violents ou non sur l'ensemble d'apprentissage et de test est récapitulé dans la Table 1.

	Objective			Subjective		
Nombre de plans	App	Test	Total	App	Test	Total
Violents	3 550	1 180	4 730	6 370	2 276	8 646
Non violents	26 476	10 065	36 541	23 656	8 969	32 625
Non identifiés	2 652	0	2 652	2 652	0	2 652
Total	32 678	11 245	43 923	32 678	11 245	43 923

Tableau 1 – Le nombre de plans annotés comme étant violents ou non sur l'ensemble d'apprentissage et de test de la collection de données de MediaEval 2013.

4.2. *Résultats*

La métrique officielle pour cette tâche est la précision moyenne sur 100 (AP@100). Nous avons comparé la performance de différents descripteurs audio et visuels : En premier temps, nous avons évalué la performance de chaque descripteur séparément. Ensuite, nous les avons opposés à la performance obtenue avec une fusion tardive en effectuant une moyenne des scores obtenus avec chaque modèle entraîné indépendamment sur chaque descripteur séparément. Enfin, nous les comparons à celles obtenues avec le descripteur joint MFCC-HOF proposé et avec la fusion du descripteur joint audio-visuel avec les deux descripteurs originaux (sacs-de-mots de MFCC et sacs-de-mots d'HOF). Nous avons rapporté dans la Figure 7 la valeur d'AP@100 pour la détection de scènes contenant de la violence objective dans 6 des 7 films de l'ensemble de test[2]. Les résultats ont montré que le descripteur audio-visuel joint est plus performant que les deux descripteurs audio (MFCC) et visuels (HOF) séparément. Le descripteur audio-visuel joint et la fusion tardive MFCC-HOF ont obtenu globalement des résultats comparables, chacun parvenant à se démarquer sur différents films. Comme supposé, le descripteur joint audio-visuel a dépassé les

2. Etant donné que le septième film (Legally blonde) ne contient aucune scène violente, nous ne l'avons pas considéré.

différents descripteurs visuel/audio pour les films contenant une vrai cohérence entre le contenu de l'image et le signal audio comme pour *Fantastic four1* et *Forrest gump*. C'est donc sur ce type de document multimédia que le descripteur a tout son intérêt.

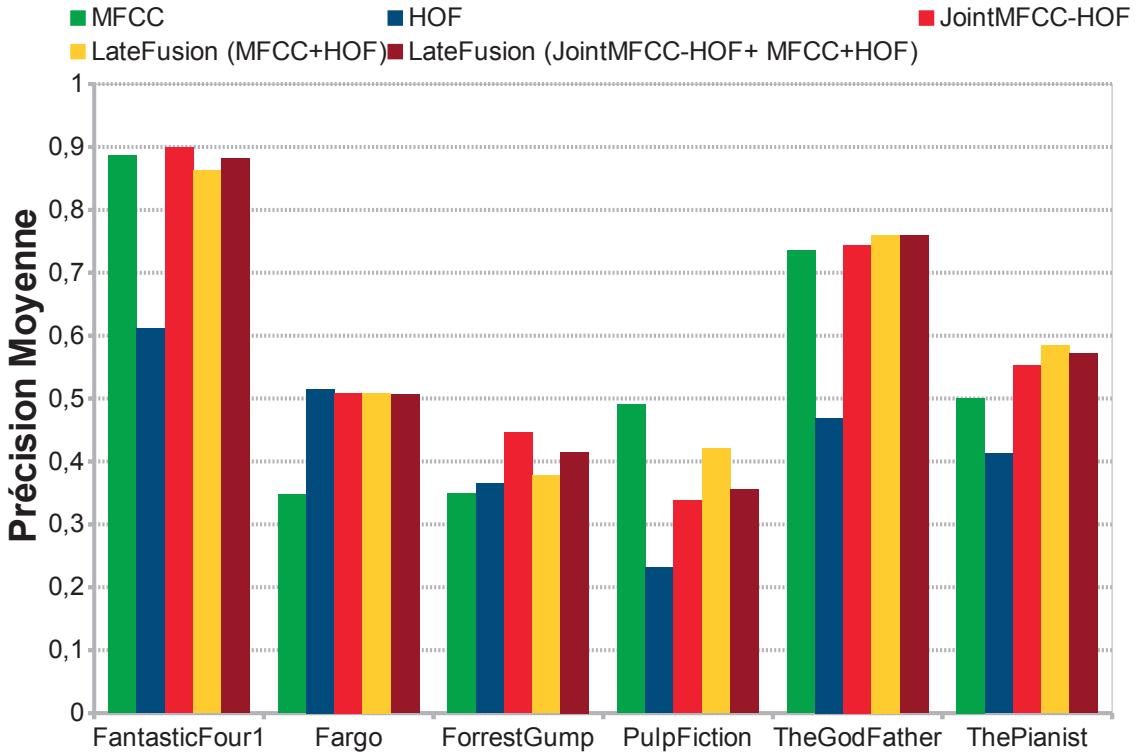

Figure 7 – L'AP@100 obtenu avec les différentes représentations en sacs-de-mots et leur fusion tardive pour la détection de scènes de violence objective sur les films de l'ensemble de test de MediaEval 2013.

Pour notre soumission officielle à la campagne d'évaluation MediaEval 2013 à la tâche de détection des scènes violentes, nous avons ajouté deux autres descripteurs (Derbas *et al.*, 2013). Le premier (OppSIFT) est basé sur le descripteur SIFT mis en place par Koen van de Sande (Van de Sande *et al.*, 2010). Alors que le deuxième est un descripteur de couleur et texture (hg104) basé sur un histogramme RGB et une transformation Gabor. Dans la Table 2, nous avons rapporté la précision moyenne à 100 (MAP@100) obtenu par notre soumission officielle, par la meilleure soumission et par la soumission médiane. La MAP@100 est la moyenne des AP@100 obtenus sur chaque film de la collection de test. Notre soumission a inclus la fusion des descripteurs MFCC, HOF, OppSIFT et hg104 avec le descripteur audio-visuel joint (Soumission avec jointAV). Cette soumission nous a classé premier sur 5 équipes participantes pour la détection de violence subjective (69%) et deuxième sur 9 équipes participantes pour la détection de violence objective (52%). En moyenne sur la violence objective et subjective, notre système a été capable de détecter les scènes violentes à 60.5%.

	Objective	Subjective	Moyenne
Meilleure Soumission	0.550	0.690	0.620
Soumission avec jointAV	0.520	0.690	0.605
Soumission Médiane	0.400	0.570	0.485

Tableau 2 – MAP@100 obtenue avec notre système à MediaEval 2013 pour la tâche de détection de scènes violentes en comparaison avec la meilleure soumission et la soumission médiane.

5. Conclusion

Dans ce papier, une nouvelle méthode a été proposée pour représenter conjointement le contenu audio-visuel dans le contexte de la détection automatique de scènes violentes. Elle exploite la corrélation entre l'information audio et l'information visuelle en construisant un dictionnaire audio-visuel joint dans le but de découvrir des motifs spécifiques audio-visuels. En comparaison avec les autres méthodes de fusion, cette méthode peut être considérée comme une fusion « pré-précoce » comme cette fusion est effectuée avant l'étape d'agrégation. Les méthodes de fusion précoces classiques, quant à elles, effectuent après l'étape d'agrégation et avant l'étape de classification.

La validation expérimentale sur de vrais films hollywoodiens (collection de données de MediaEval 2013) a montré que la fusion audio-visuelle jointe donne des résultats comparables à ceux obtenus avec la fusion tardive. Plusieurs pistes pourront être considérées dans le futur, comme l'application de cette représentation conjointe à d'autres types de concepts dynamiques (autres que la violence). L'utilisation du descripteur joint audio-visuel à une échelle plus petite que celle du plan en entier est également envisagée vu que l'utilité de la corrélation peut être mieux capturée avec une localisation temporelle plus étroite. Pour finir, ce travail pourrait être étendu pour intégrer plus que deux descripteurs originaux (MFCC et STIP-HOF).

Remerciements

Ce travail a été réalisé partiellement dans le cadre du programme Quaero et du projet Camomile, financés respectivement par OSEO (l'agence française pour l'innovation) et ANR (l'agence nationale française pour la recherche).

6. Bibliographie

Atrey P. K., Hossain M. A., El Saddik A., Kankanhalli M. S., « Multimodal fusion for multimedia analysis : a survey », *Multimedia systems*, vol. 16, n° 6, p. 345-379, 2010.

Ayache S., Quénot G., Gensel J., « Classifier fusion for SVM-based multimedia semantic indexing », *Advances in Information Retrieval*, Springer, p. 494-504, 2007.

Beal M. J., Jojic N., Attias H., « A graphical model for audiovisual object tracking », *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, n° 7, p. 828-836, 2003.

Bermejo Nievas E., Deniz Suarez O., Bueno García G., Sukthankar R., « Violence Detection in Video Using Computer Vision Techniques », *Computer Analysis of Images and Patterns*, Springer Berlin Heidelberg, p. 332-339, 2011.

Cristani M., Bicego M., Murino V., « Audio-Visual Event Recognition in Surveillance Video Sequences », *Multimedia, IEEE Transactions on*, vol. 9, n° 2, p. 257-267, 2007.

Datta A., Shah M., da Vitoria Lobo N., « Person-on-person violence detection in video data », *Pattern Recognition*, vol. 1, p. 433-438 vol.1, 2002.

de Souza F., Chávez G., do Valle E., de A Araujo A., « Violence Detection in Video Using Spatio-Temporal Features », *Proceedings of the 2010 23rd SIBGRAPI Conference on Graphics, Patterns and Images*, n° 7, Washington, DC, USA, p. 224-230, August 30-Septembre 3, 2010.

Demarty C.-H., Penet C., Schedl M., Ionescu B., Quang V. L., Jiang Y.-G., « The MediaEval 2013 Affect Task : Violent Scenes Detection », *MediaEval Workshop*, Barcelona, Spain, October 18-19, 2013.

Derbas N., Safadi B., Quénot G., « LIG at MediaEval 2013 Affect Task : Use of a Generic Method and Joint Audio-Visual Words », *MediaEval Workshop*, Barcelona, Spain, October 18-19, 2013.

Derbas N., Thollard F., Safadi B., Quénot G., « LIG at MediaEval 2012 Affect Task : Use of a Generic Method », *MediaEval Workshop*, Pisa, Italy, October 4-5, 2012.

Giannakopoulos T., Kosmopoulos D. I., Aristidou A., Theodoridis S., « Violence Content Classification Using Audio Features », *SETN*, p. 502-507, 2006.

Giannakopoulos T., Makris A., Kosmopoulos D., Perantonis S., Theodoridis S., « Audio-Visual Fusion for Detecting Violent Scenes in Videos », *Artificial Intelligence : Theories, Models and Applications*, Springer Berlin Heidelberg, p. 91-100, 2010.

Gong Y., Wang W., Jiang S., Huang Q., Gao W., « Detecting Violent Scenes in Movies by Auditory and Visual Cues », *Advances in Multimedia Information Processing - PCM 2008*, Springer Berlin Heidelberg, p. 317-326, 2008.

Jiang W., Loui A. C., « Audio-visual grouplet : temporal audio-visual interactions for general video concept classification », *ACM Multimedia*, p. 123-132, 2011.

Laptev I., « On Space-Time Interest Points », *Int. J. Comput. Vision*, vol. 64, n° 2-3, p. 107-123, September, 2005.

Lin J., Wang W., « Weakly-Supervised Violence Detection in Movies with Audio and Video Based Co-training », *Advances in Multimedia Information Processing - PCM 2009*, Springer Berlin Heidelberg, p. 930-935, 2009.

Mühling M., Ewerth R., Zhou J., Freisleben B., « Multimodal video concept detection via bag of auditory words and multiple kernel learning », *Advances in Multimedia Modeling*, Springer, p. 40-50, 2012.

Penet C., Demarty C.-H., Gravier G., Gros P., « Audio event detection in movies using multiple audio words and contextual Bayesian networks », *Workshop on Content-Based Multimedia Indexing*, p. 17-22, 2013.

Safadi B., Quénot G., « Evaluations of multi-learner approaches for concept indexing in video documents », *Adaptivity, Personalization and Fusion of Heterogeneous Information*, RIAO '10, p. 88-91, 2010.

Snoek C. G. M., Worring M., Smeulders A. W. M., « Early Versus Late Fusion in Semantic Video Analysis », *ACM International Conference on Multimedia*, p. 399—402, 2005.

Van de Sande K. E., Gevers T., Snoek C. G., « Evaluating color descriptors for object and scene recognition », *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, n° 9, p. 1582-1596, 2010.

Ye G., Jhuo I.-H., Liu D., Jiang Y.-G., Lee D., Chang S.-F., « Joint Audio-Visual Bi-Modal Codewords for Video Event Detection », *ACM International Conference on Multimedia Retrieval (ICMR)*, Hong Kong, June 5-8, 2012.